Technical Report OSU-CISRC-9/14-TR16
Department of Computer Science and Engineering
The Ohio State University
Columbus, OH 43210-1277

Ftpsite: **ftp.cse.ohio-state.edu**
Login: **anonymous**
Directory: **pub/tech-report/2014**
File: **TR16.pdf**
Website: **http://www.cse.ohio-state.edu/research/techReport.shtml**

Noise Perturbation Improves Supervised Speech Separation

Jitong Chen
Department of Computer Science and Engineering
The Ohio State University, Columbus, OH 43210, USA
chenjit@cse.ohio-state.edu

Yuxuan Wang
Department of Computer Science and Engineering
The Ohio State University, Columbus, OH 43210, USA
wangyuxu@cse.ohio-state.edu

DeLiang Wang
Department of Computer Science and Engineering & Center for Cognitive and Brain Sciences
The Ohio State University, Columbus, OH 43210, USA
dwang@cse.ohio-state.edu

Abstract – Speech separation can be treated as a mask estimation problem where interference-dominant portions are masked in a time-frequency representation of noisy speech. In supervised speech separation, a classifier is typically trained on a mixture set of speech and noise. It is important to efficiently utilize limited training data to make the classifier generalize well. When target speech is severely interfered by a nonstationary noise, a classifier tends to mistake noise patterns for speech patterns. Expansion of a noise through proper perturbation during training helps to expose the classifier to a broader variety of noisy conditions, and hence may improve separation performance. In this study, we examine the effects of three noise perturbations on supervised speech separation: noise rate, vocal tract length, and frequency perturbation at low signal-to-noise ratios (SNRs). We evaluate speech separation performance in terms of classification accuracy, hit minus false-alarm rate and short-time objective intelligibility (STOI). The experimental results show that frequency perturbation is the best among the three perturbations in terms of improved speech separation. In particular, we find that

frequency perturbation is effective in reducing the error of misclassifying a noise pattern as a speech pattern.

Index Terms – Speech separation, supervised learning, noise perturbation.

1 Introduction

Speech separation is a task of separating target speech from noise interference. The task has a wide range of applications such as hearing aid design and robust automatic speech recognition (ASR). Monaural speech separation is proven to be very challenging as it only uses single-microphone recordings, especially in low SNR conditions. One way of dealing with this problem is to apply speech enhancement [7] [8] [14] on a noisy signal, where certain assumptions are made regarding general statistics of the background noise. The speech enhancement approach is usually limited to relatively stationary noises. Looking at the problem from another perspective, computational auditory scene analysis (CASA) [24], which is inspired by psychoacoustic research in auditory scene analysis (ASA) [2], exploits perceptual principles to speech separation.

In CASA, interference can be reduced by applying masking on a time-frequency (T-F) representation of noisy speech. An ideal mask suppresses noise-dominant T-F units and keeps the speech-dominant T-F units. Therefore, speech separation can be treated as a mask estimation problem where supervised learning is employed to construct the mapping from acoustic features to a mask. A binary decision on each T-F unit leads to an estimate of the ideal binary mask (IBM), which is defined as follows.

$$\mathrm{IBM}(t,f) = \begin{cases} 1, & \text{if } \mathrm{SNR}(t,f) > \mathrm{LC} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where t denotes time and f frequency. The IBM assigns the value 1 to a T-F unit if its SNR exceeds a local criterion (LC), and 0 otherwise. Therefore, speech separation is translated into a binary classification problem. Recent studies show IBM separation improves speech intelligibility in noise for both normal-hearing and hearing-impaired listeners [3] [18] [25] [1]. Alternatively, a soft decision on each T-F unit leads to an estimate of the ideal ratio mask (IRM). The IRM is defined below [21].

$$\mathrm{IRM}(t,f) = \left(\frac{10^{(\mathrm{SNR}(t,f)/10)}}{10^{(\mathrm{SNR}(t,f)/10)} + 1}\right)^{\beta} \tag{2}$$

where β is a tunable parameter. A recent study has shown that $\beta = 0.5$ is a good choice for the IRM [27]. In this case, mask estimation becomes a regression problem where the target is the IRM. Ratio masking is shown to lead to slightly better objective intelligibility results than binary masking [27]. In this study, we use the IRM with $\beta = 0.5$ as the learning target.

Supervised speech separation is a data-driven method where one expects a mask estimator to generalize from limited training data. However, training data only partially captures the true data distribution, thus a mask estimator can overfit training data and do a poor job in unseen scenarios. In supervised speech separation, a training set is typically created by mixing clean speech and noise. When we train and test on a nonstationary noise such as a cafeteria

noise, there can be considerable mismatch between training noise segments and test noise segments, especially when the noise resource used for training is restricted. Similar problems can be seen in other supervised learning tasks such as image classification where the mismatch of training images and test images poses a great challenge. In image classification, a common practice is to transform training images using distortions such as rotation, translation and scaling, in order to expand the training set and improve generalization of a classifier [17] [4]. We conjecture that supervised speech separation can also benefit from training data augmentation.

In this study, we aim at expanding the noise resource using noise perturbation to improve supervised speech separation. We treat noise expansion as a way to prevent a mask estimator from overfitting the training data. A recent study has shown speech perturbation improves ASR [15]. However, our study perturbs noise instead of speech since we focus on separating target speech from highly nonstationary noises where the mismatch among noise segments is the major problem.

This paper is organized as follows. Section 2 describes the system used for mask estimation. Noise perturbations are covered in section 3. We present experimental results in section 4. Section 5 concludes the paper.

2 System Overview

To evaluate the effects of noise perturbation, we use a fixed system for mask estimation and compare the quality of estimated masks as well as the resynthesized speech that are derived from the masked T-F representations of noisy speech. While comparison between an estimated mask and an ideal mask reveals the spectrotemporal distribution of estimation errors, resythesized speech can be directly compared to clean speech. As mentioned in Section 1, we use the IRM as the target of supervised learning. The IRM is computed from the 64-channel cochleagrams of premixed clean speech and noise. The cochleagram is a time-frequency representation of a signal [24]. We use a 20 ms window and a 10 ms window shift to compute cochleagram in this study.

We perform IRM estimation using a deep neural network (DNN) and a set of acoustic features. Recent studies have shown that DNN is a strong classifier for ASR [19] and speech separation [28]. As shown in Fig. 1, acoustic features are extracted from a mixture sampled at 16 kHz, and then sent to a DNN for mask prediction. To incorporate temporal context and obtain smooth mask estimation, we use 5 frames of features to estimate 5 frames of the IRM [27]. Since each frame of the mask is estimated 5 times, we take the average of the 5 estimates.

The acoustic features we extract from mixtures are a complementary feature set (AMS + RASTAPLP + MFCC) [26] combined with gammatone filterbank (GFB) features. To compute 15-D AMS, we derive 15 modulation spectrum amplitudes from the decimated envelope of an input signal [16]. 13-D RASTAPLP is derived by applying linear prediction analysis on the

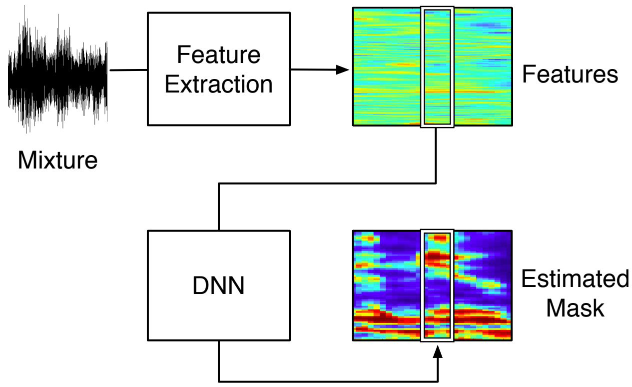


Figure 1: Diagram of the proposed system.

RASTA-filtered bark-scale power spectrum of an input signal [11]. We follow a standard procedure to compute 31-D MFCC. To derive GFB features, an input signal is passed to a 64-channel gammatone filterbank, the response signals are decimated to 100 Hz to form 64-D GFB features.

We use classification accuracy, hit minus false-alarm (HIT−FA) rate and short-time objective intelligibility (STOI) score [22] as three criteria for measuring the quality of the estimated IRM. Since the first two criteria are defined for binary masks, we calculate them by binarizing a ratio mask to a binary one. In this study, we follow Equation 3 and Equation 1.

$$\mathrm{SNR}(t,f) = 10log_{10}(\frac{\mathrm{IRM}(t,f)^2}{1-\mathrm{IRM}(t,f)^2}) \tag{3}$$

During the mask conversion, the LC is set to be 5 dB lower than the SNR of a given mixture. The three criteria evaluate the estimated IRM from three different perspectives. Classification accuracy computes the percentage of correctly labeled T-F units in a binary mask. In HIT−FA, HIT refers to the percentage of correctly classified target-dominant T-F units and FA refers to the percentage of wrongly classified interference-dominant T-F units. HIT−FA rate is well correlated with human speech intelligibility [16]. In addition, STOI is computed by comparing the the short-time envelopes of clean speech and resynthesized speech obtained from IRM masking, and it is a standard objective metric of speech intelligibility [22].

3 Noise perturbation

The goal of noise perturbation is to expand noise segments to cover unseen scenarios so that the overfitting problem is mitigated in supervised speech separation. A recent study has found that three perturbations on speech samples improve ASR performance [15]. These perturbations were used to expand the speech samples by spectral perturbation. The three perturbations are introduced below. Unlike this study, we perturb noise samples instead of

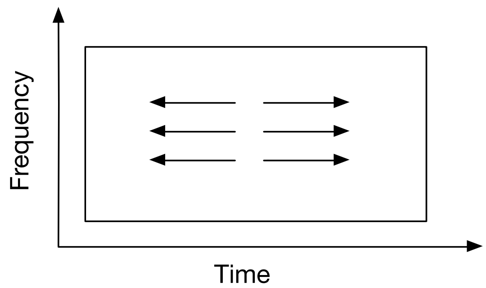


Figure 2: Illustration of noise rate perturbation.

perturbing speech samples, as we are dealing with highly nonstationary noises.

3.1 Noise Rate (NR) Perturbation

Speech rate perturbation, a way of speeding up or slow down speech, is used to expand training utterances during the training of an ASR system. In our study, we extend the method to vary the rate of nonstationary noises. We increase or decrease noise rate by factor γ. When a noise rate is being perturbed, the value of γ is randomly selected from an interval $[\gamma_{min}, 2-\gamma_{min}]$. The effect of NR perturbation on a spectrogram is shown in Fig. 2.

3.2 Vocal Tract Length (VTL) Perturbation

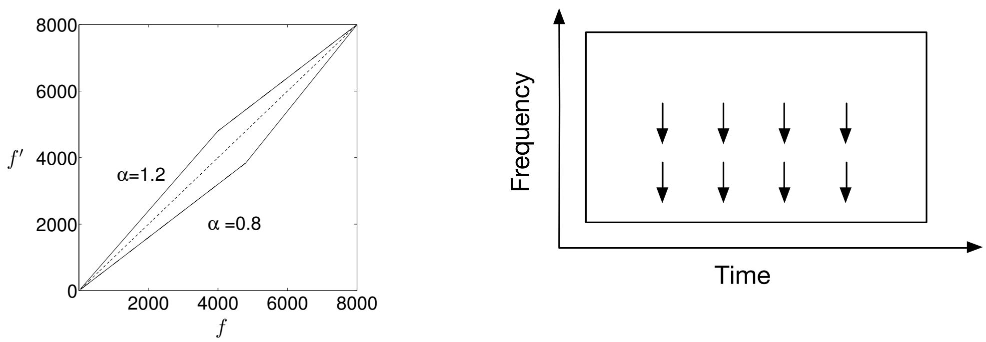
VTL perturbation has been used in ASR to cover the variation of vocal tract length among speakers. A recent study suggests that VTL perturbation improves ASR performance [13]. VTL perturbation essentially compresses or stretches the medium and low frequency components of an input signal. We use VTL perturbation as a method of perturbing a noise segment. Specifically, we follow the algorithm in [13] to perturb noise signals:

$$f' = \begin{cases} f\alpha, & \text{if } f \leq F_{hi}\frac{min(\alpha,1)}{\alpha} \\ \frac{S}{2} - \frac{\frac{S}{2}-F_{hi}min(\alpha,1)}{\frac{S}{2}-F_{hi}\frac{min(\alpha,1)}{\alpha}}(\frac{S}{2}-f), & \text{otherwise} \end{cases} \tag{4}$$

where α is the wrapping factor, S is the sampling rate, and F_{hi} controls the cutoff frequency. Fig. 3(a) shows how VTL perturbation compresses or stretches a portion of a spectrogram. The effect of VTL perturbation is visualized in Fig. 3(b).

3.3 Frequency Perturbation

When frequency perturbation is applied, frequency bands of a spectrogram are randomly shifted upward or downward. We use the method described in [15] to randomly perturb noise samples. Frequency perturbation takes three steps. First, we randomly assign a value to each



(a) IBM (b) TBM

Figure 3: (a) Mapping function for vocal tract length perturbation. The frequencies below a cutoff are stretched if $\alpha > 1$, and compressed if $\alpha < 1$. (b) Illustration of vocal tract length perturbation. The medium and low frequencies are compressed in this case.

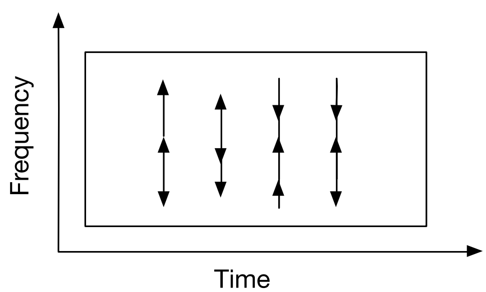


Figure 4: Illustration of frequency perturbation.

T-F unit, which is drawn from a uniform distribution.

$$r(f,t) \sim U(-1,1) \tag{5}$$

Then we derive the perturbation factor $\delta(f,t)$ by averaging the assigned values of neighboring time-frequency units. This averaging step avoids large oscillations in spectrogram.

$$\delta(f,t) = \frac{\lambda}{(2p+1)(2q+1)} \sum_{f'=f-p}^{f+p} \sum_{t'=t-q}^{t+q} r(f',t') \tag{6}$$

where p and q control the smoothness of the perturbation, and λ controls the magnitude of the perturbation. These tunable parameters are decided experimentally. Finally the spectrogram is perturbed as follows.

$$\tilde{S}(f,t) = S(f+\delta(f,t),t) \tag{7}$$

where $S(f,t)$ represents the original spectrogram and $\tilde{S}(f,t)$ is the perturbed spectrogram. Interpolation between neighboring frequencies is used when $\delta(f,t)$ is not an integer. The effect of frequency perturbation is visualized in Fig. 4.

4 Experimental Results

4.1 Experimental Setup

We use the IEEE corpus recorded by a male speaker [12] and six nonstationary noises from the DEMAND corpus [23] to create mixtures. All signals are sampled at 16 KHz. Note that all recordings of the DEMAND corpus are made with a 16-channel microphone array, we use only one channel of the recordings since this study is on monaural speech separation.

The DEMAND corpus has six categories of noises. We choose one noise from each category to represent distinct environments. The six nonstationary noises, each is five-minute long, are described as follows.

1. The "Street" category:
   The SCAFE noise, recorded in the terrace of a cafe at a public square.
2. The "Domestic" category:
   The DLIVING noise, recorded inside a living room.
3. The "Office" category:
   The OMEETING noise, recorded in a meeting room.
4. The "Public" category:
   The PCAFETER noise, recorded in a busy office cafeteria.
5. The "Nature" category:
   The NPARK noise, recorded in a well visited city park.
6. The "Transportation" category:
   The TMETRO noise, recorded in a subway.

To create a mixture, we mix one IEEE sentence and one noise type at -5 dB SNR. This low SNR is selected with the goal of improving speech intelligibility in mind where there is not much to improve at higher SNRs [10]. The training set uses 600 IEEE sentences and randomly selected segments from the first two minutes of a noise, while the test set uses another 120 IEEE sentences and randomly selected segments from the second two minutes of a noises. Therefore, the test set has different sentences and different noise segments from the training set. We create 50 mixtures for each training sentence by mixing it with 50 randomly selected segments from a given noise, which results in a training set containing 600×50 mixtures. The test set includes 120 mixtures. We train and test using the same noise type and SNR condition.

To perturb a noise segment, we first apply short-time Fourier transform (STFT) to derive noise spectrogram, where a frame length of 20 ms and a frame shift of 10 ms are used. Then we perturb the spectrogram and derive a new noise segment. To evaluate the three noise perturbations, we create five different training sets, each consists of 600×50 mixtures. We

train a mask estimator for each training set and evaluate on a fixed test set (i.e. the 120 mixtures created from the original noises). The five training sets are described as follows.

1. Original Noise: All mixtures are created using original noises.
2. NR Perturbation: Half of the mixtures are created from NR perturbed noises, and the other half are from original noises.
3. VTL Perturbation: Half of the mixtures are created from VTL perturbed noises, and the other half are from original noises.
4. Frequency Perturbation: Half of the mixtures are created from frequency perturbed noises, and the other half are from original noises.
5. Combined: Half of the mixtures are created from applying three perturbations altogether, and the other half are from original noises.

As already mentioned, we extract a set of four complementary features (AMS + RASTA-PLP + MFCC + GFB) from mixtures. Delta features are appended to the feature set. A four-hidden-layer DNN is employed to learn the mapping from acoustic features to the IRM. Each hidden layer of the DNN has 1024 rectified linear units [20]. Dropout [5] and adaptive stochastic gradient descent [6] are used to train the DNN.

4.2 Parameters of Noise Perturbation

In this section, three sets of experiments are carried out to explore the parameters used in the three perturbations to get the best performance. To facilitate parameter selection, we create five smaller training sets, following the same configuration in Section 4.1 except that we use 480 IEEE clean sentences to create 480×20 training mixtures. Another 120 IEEE sentences (different than the test ones in Section 4.1) are used to create 120 test mixtures only for the purpose of choosing parameter values (i.e. a development set). The speech separation performance is evaluated in term of STOI score.

In NR perturbation, the only adjustable parameter is the rate γ. We can slow down a noise by setting $\gamma < 1$, or speed it up using $\gamma > 1$. To capture various noise rates, we randomly draw γ from an interval $[\gamma_{min}, 2 - \gamma_{min}]$. We evaluate various intervals in term of speech separation performance. As shown in Fig. 5, the interval $[0.1, 1.9]$ (i.e. $\gamma_{min} = 0.1$) gives the best performance for six noises.

In VTL perturbation, there are two parameters: F_{hi} controls cutoff frequency and α the warping factor. F_{hi} is set to 4800 to roughly cover the frequency range of speech formants. We randomly draw α from an interval $[\alpha_{min}, 2 - \alpha_{min}]$ to systematically stretch or shrink the frequencies below the cutoff frequency. Fig. 6 shows the effects of different intervals on speech separation performance. The interval of $[0.3, 1.7]$ (i.e. $\alpha_{min} = 0.3$) leads to the best result for the majority of the noise types.

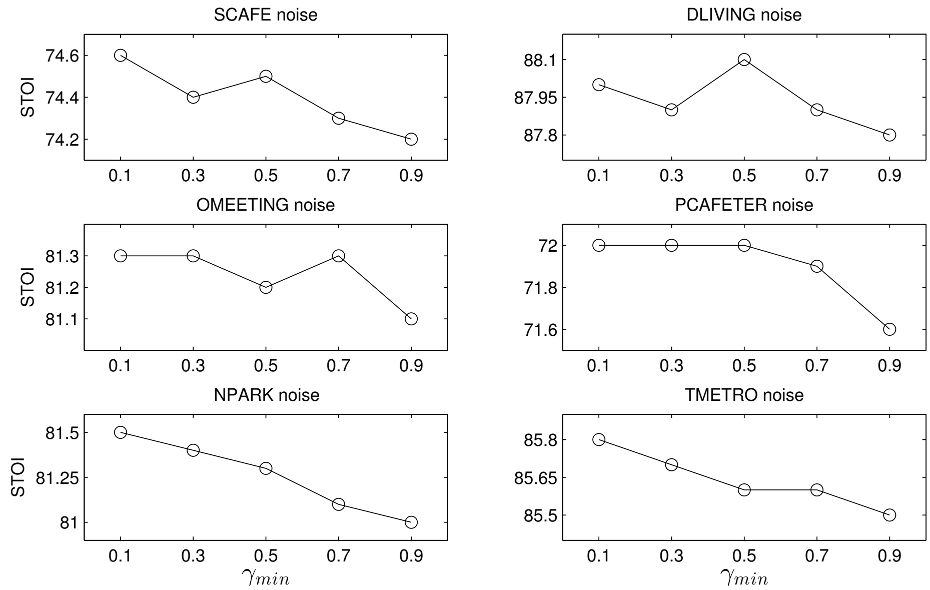


Figure 5: The effect of the minimum noise rate γ_{min} for NR perturbation.

In frequency perturbation, a 161-band spectrogram derived from a noise segment is perturbed using the algorithm described in Section 3.3. We set $p = 50$ and $q = 100$ to avoid dramatic perturbation along time and frequency axes. We experiment with different perturbation intensity λ. As shown in Fig. 7, $\lambda = 1000$ achieves the best performance for the majority of the noise types.

4.3 Evaluation Results and Comparisons

We evaluate the three perturbations with the parameter values selected in Section 4.2 and the five large training sets described in Section 4.1. The effects of noise perturbations on speech separation are shown in Table 1, Table 2 and Table 3, in terms of classification accuracy, HIT−FA rate and STOI score respectively. The results indicate that all three perturbations lead to better speech separation than the baseline where only the original noises are used. Frequency perturbation performs better than the other two perturbations. Compared to only using the original noises, the frequency perturbed training set on average increases classification accuracy, HIT−FA rate and STOI score by 8%, 11% and 3%, respectively. This indicates that noise perturbation is an effective technique for improving speech separation results. Combining three perturbations, however, does not lead to further improvement over frequency perturbation.

A closer look at Table 2 reveals that the contribution of frequency perturbation lies mainly in the large reduction in FA rate. This means that the problem of misclassifying noise-dominant T-F units as speech-dominant is mitigated. This effect can be illustrated by visualizing the masks estimated from the different training sets and the ground truth mask in

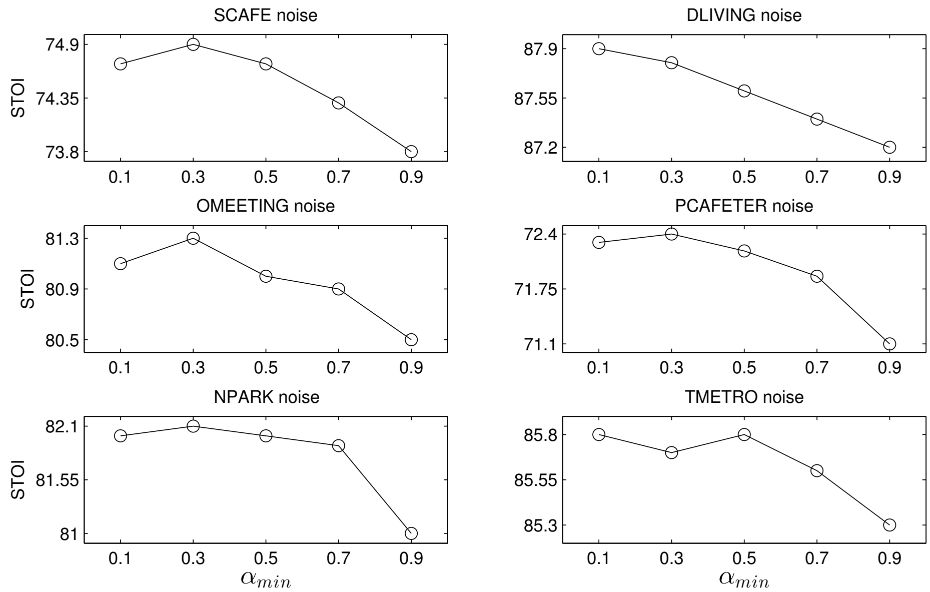


Figure 6: The effect of the minimum wrapping factor α_{min} for VTL perturbation.

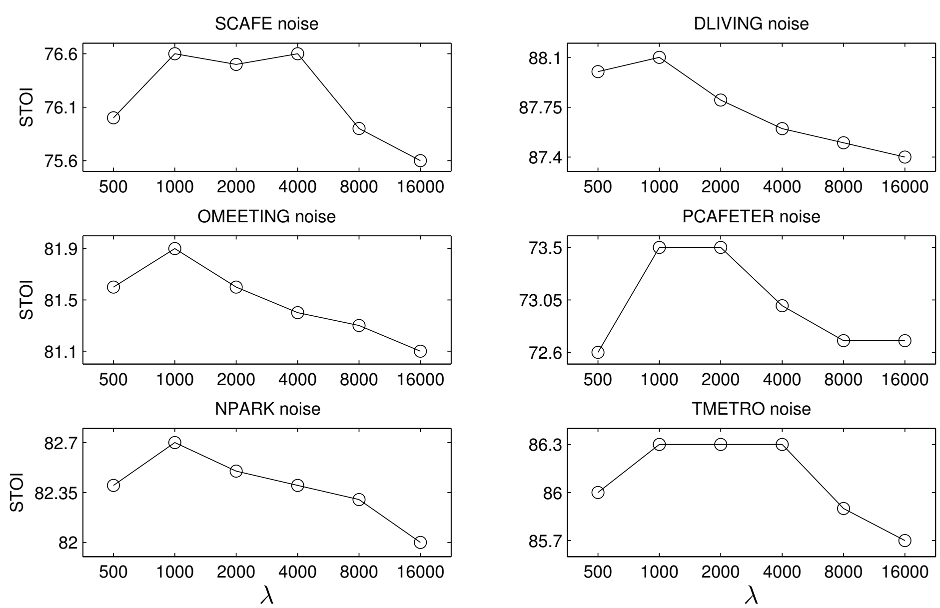


Figure 7: The effect of the perturbation intensity λ for frequency perturbation.

Table 1: Classification accuracy (in %) for six noises at -5 dB

Perturbation \ Noise	SCAFE	DLIVING	OMEETING	PCAFETER	NPARK	TMETRO	Average
Original Noise	73.0	84.0	80.0	70.3	82.7	80.3	78.4
NR Perturbation	80.2	88.5	85.3	77.9	88.5	85.1	84.2
VTL Perturbation	80.1	87.7	84.9	77.8	89.2	85.5	84.2
Frequency Perturbation	84.4	88.6	86.7	80.6	90.0	86.7	86.2
Combined	81.8	88.0	86.1	78.9	89.6	86.6	85.2

Table 2: HIT−FA rate (in %) for six noises at -5 dB, where FA is shown in parentheses.

Perturbation \ Noise	SCAFE	DLIVING	OMEETING	PCAFETER	NPARK	TMETRO	Average
Original Noise	55 (37)	70 (23)	65 (28)	50 (40)	69 (22)	63 (32)	62 (30)
NR perturbation	64 (24)	77 (15)	72 (18)	60 (26)	77 (12)	72 (21)	70 (19)
VTL Perturbation	64 (24)	76 (16)	71 (19)	60 (27)	78 (10)	72 (21)	70 (20)
Frequency Perturbation	69 (17)	77 (14)	74 (15)	63 (21)	79 (9)	74 (18)	73 (16)
Combined	67 (21)	77 (15)	73 (16)	61 (25)	78 (10)	74 (18)	72 (18)

Table 3: STOI (in %) of separated speech for six noises at -5 dB, where STOI of unprocessed mixtures is shown in parentheses.

Perturbation \ Noise	SCAFE	DLIVING	OMEETING	PCAFETER	NPARK	TMETRO	Average
Original Noise	73.7 (64.1)	87.5 (79.3)	80.0 (67.8)	71.4 (62.5)	80.2 (67.7)	85.9 (77.5)	79.8 (69.8)
NR perturbation	76.5 (64.1)	89.2 (79.3)	82.5 (67.8)	74.1 (62.5)	83.2 (67.7)	87.4 (77.5)	82.1 (69.8)
VTL Perturbation	76.1 (64.1)	88.7 (79.3)	82.2 (67.8)	74.0 (62.5)	83.6 (67.7)	87.2 (77.5)	82.0 (69.8)
Frequency Perturbation	78.2 (64.1)	89.1 (79.3)	83.3 (67.8)	75.1 (62.5)	84.1 (67.7)	87.8 (77.5)	82.9 (69.8)
Combined	77.0 (64.1)	88.6 (79.3)	82.7 (67.8)	74.7 (62.5)	83.8 (67.7)	87.6 (77.5)	82.4 (69.8)

Fig. 8 (e.g. around frame 150). When the mask estimator is trained with the original noises, it mistakenly retains the regions where target speech is not present, which can be seen by comparing the top and bottom plots of Fig. 8. Applying frequency perturbation to noises essentially exposes the mask estimator to more noise patterns and results in a more accurate mask estimator, which is shown in the middle plot of Fig. 8.

In addition, we show HIT−FA rate for voiced and unvoiced intervals in Table 4 and Table 5 respectively. We find that frequency perturbation is effective for both voiced and unvoiced intervals.

While classification accuracy and HIT−FA rate evaluate the estimated binary masks, STOI directly compares clean speech and the resynthesized speech. As shown in Table 3, frequency perturbation yields higher average STOI scores than using original noises with no perturbation and NR and VTL perturbations.

Finally, to evaluate the effectiveness of frequency perturbation at other SNRs, we carry out additional experiments at -10 dB and 0 dB input SNRs, where we use the same parameter

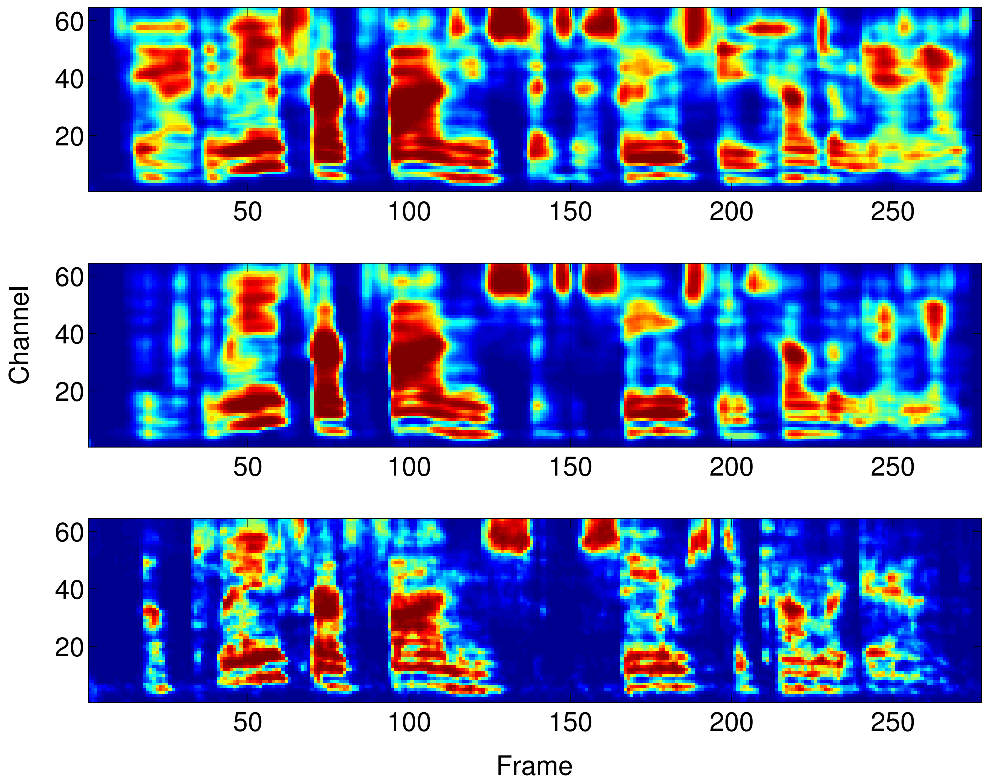


Figure 8: Mask comparisons. The top shows a ratio mask obtained from training on original noises, the middle shows a mask obtained from training on frequency perturbed noise, and the bottom shows the IRM.

Table 4: HIT−FA rate (in %) during voiced intervals, where FA is shown in parentheses.

Noise / **Perturbation**	SCAFE	DLIVING	OMEETING	PCAFETER	NPARK	TMETRO	Average
Original Noise	50 (44)	70 (26)	62 (33)	48 (45)	71 (24)	55 (42)	59 (36)
NR perturbation	60 (32)	75 (21)	69 (24)	57 (33)	79 (15)	63 (33)	67 (26)
VTL Perturbation	62 (30)	75 (21)	70 (24)	60 (31)	80 (13)	65 (31)	69 (25)
Frequency Perturbation	66 (24)	76 (20)	72 (21)	62 (27)	80 (13)	67 (29)	70 (22)
Combined	65 (27)	76 (20)	72 (21)	61 (30)	80 (13)	68 (28)	70 (23)

Table 5: HIT−FA rate (in %) during unvoiced intervals, where FA is shown in parentheses.

Noise / **Perturbation**	SCAFE	DLIVING	OMEETING	PCAFETER	NPARK	TMETRO	Average
Original Noise	48 (33)	61 (22)	59 (25)	41 (36)	57 (20)	61 (27)	54 (27)
NR perturbation	54 (20)	70 (11)	64 (15)	48 (22)	62 (9)	68 (16)	61 (16)
VTL Perturbation	52 (21)	68 (13)	64 (15)	45 (24)	62 (8)	68 (16)	60 (16)
Frequency Perturbation	59 (12)	68 (11)	66 (11)	48 (18)	62 (6)	70 (13)	62 (12)
Combined	55 (18)	68 (12)	64 (13)	46 (22)	62 (8)	69 (14)	61 (14)

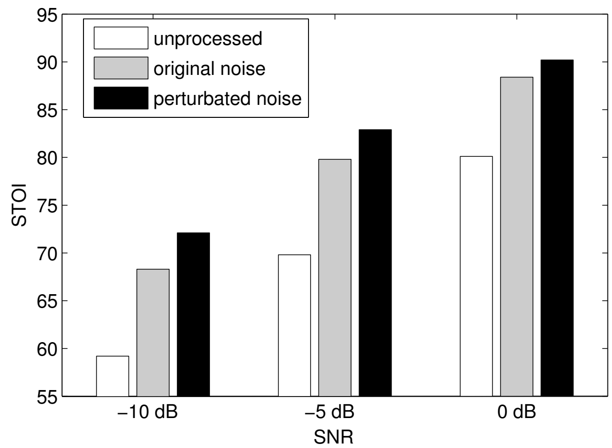


Figure 9: The effect of frequency perturbation in three SNR conditions. The average STOI scores (in %) across six noises are shown for unprocessed speech, separated speech by training on original noises, and separated speech by training on frequency perturbed noises.

values as for -5 dB SNR. Fig. 9 shows frequency perturbation improves speech separation in terms of STOI in each SNR condition. Also, we find that frequency perturbation remains the most effective among the three perturbations at -10 dB and 0 dB SNR.

5 Concluding Remarks

In this study, we have explored the effects of noise perturbation on supervised monaural speech separation at low SNR levels. As a training set is usually created from limited speech and noise resources, a classifier likely overfits the training set and makes poor predictions on a test set, especially when background noise is highly nonstationary. We suggest to expand limited noise resources by noise perturbation.

We have evaluated three noise perturbations with six nonstationary noises recorded from daily life for speech separation. The three are noise rate, VTL, and frequency perturbations. When a DNN is trained on a data set which utilizes perturbed noises, the quality of the estimated ratio mask is improved as the classifier has been exposed to more scenarios of noise interference. In contrast, a mask estimator learned from a training set that only uses original noises tends to make more false alarm errors (i.e. higher FA rate), which is detrimental to speech intelligibility [29]. The experimental results show that frequency perturbation, which randomly perturbs the noise spectrogram along frequency, almost uniformly gives the best speech separation results among the three perturbations examined in this study in terms of classification accuracy, HIT−FA rate and STOI score.

Finally, this study adds another technique to deal with the generalization problem in supervised speech separation. Previous studies use model adaptation [9] and extensive training [28] to deal with the mismatch of SNR conditions, noises and speakers between training and test-

ing. Our study aims at situations with limited training noises, and provides an effective data augmentation method that improves generalization in nonstationary environments.

Acknowledgments

This research was supported in part by an AFOSR grant (FA9550-12-1-0130), an NIDCD grant (R01 DC012048) and the Ohio Supercomputer Center.

References

[1] M. Ahmadi, V. L. Gross, and D. G. Sinex, "Perceptual learning for speech in noise after application of binary time-frequency masks," *J. Acoust. Soc. Am.*, vol. 133, pp. 1687–1692, 2013.

[2] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound.* Cambridge MA: MIT Press, 1994.

[3] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. L. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *J. Acoust. Soc. Am.*, vol. 120, pp. 4007–4018, 2006.

[4] D. Ciresan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *Proc. CVPR*, 2012, pp. 3642–3649.

[5] G. E. Dahl, T. N. Sainath, and G. E. Hinton, "Improving deep neural networks for LVCSR using rectified linear units and dropout," in *Proc. ICASSP*, 2013, pp. 8609–8613.

[6] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *The Journal of Machine Learning Research*, vol. 12, pp. 2121–2159, 2011.

[7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Sig. Process.*, vol. 32, pp. 1109–1121, 1984.

[8] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, pp. 1741–1752, 2007.

[9] K. Han and D. Wang, "Towards generalizing classification based speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, pp. 168–177, 2013.

[10] E. W. Healy, S. E. Yoho, Y. Wang, and D. L. Wang, "An algorithm to improve speech recognition in noise for hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 134, pp. 3029–3038, 2013.

[11] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech, Audio Process.*, vol. 2, pp. 578–589, 1994.

[12] IEEE, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. 17, pp. 225–246, 1969.

[13] N. Jaitly and G. E. Hinton, "Vocal Tract Length Perturbation (VTLP) improves speech recognition," in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Lang. Process.*, 2013.

[14] J. Jensen and R. C. Hendriks, "Spectral magnitude minimum mean-square error estimation using binary and continuous gain functions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, pp. 92–102, 2012.

[15] N. Kanda, R. Takeda, and Y. Obuchi, "Elastic spectral distortion for low resource speech recognition with deep neural networks," in *Proc. ASRU*, 2013, pp. 309–314.

[16] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *J. Acoust. Soc. Am.*, vol. 126, pp. 1486–1494, 2009.

[17] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. of the IEEE*, vol. 86, pp. 2278–2324, 1998.

[18] N. Li and P. C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *J. Acoust. Soc. Am.*, vol. 123, pp. 1673–1682, 2008.

[19] A. Mohamed, G. E. Dahl, and G. E. Hinton, "Acoustic modeling using deep belief networks," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, pp. 14–22, 2012.

[20] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. ICML*, 2010, pp. 807–814.

[21] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. ICASSP*, 2013, pp. 7092–7096.

[22] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, pp. 2125–2136, 2011.

[23] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings," *J. Acoust. Soc. Am.*, vol. 133, p. 3591, 2013.

[24] D. L. Wang and G. J. Brown, Eds., *Computational auditory scene analysis: Principles, algorithms and applications.* Hoboken NJ: Wiley-IEEE Press, 2006.

[25] D. L. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, "Speech intelligibility in background noise with ideal binary time-frequency masking," *J. Acoust. Soc. Am.*, vol. 125, pp. 2336–2347, 2009.

[26] Y. Wang, K. Han, and D. L. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, pp. 270–279, 2013.

[27] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, in press, 2014.

[28] Y. Wang and D. L. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, pp. 1381–1390, 2013.

[29] C. Yu, K. K. Wójcicki, P. C. Loizou, J. H. Hansen, and M. T. Johnson, "Evaluation of the importance of time-frequency contributions to speech intelligibility in noise," *J. Acoust. Soc. Am.*, vol. 135, pp. 3007–3016, 2014.